A Bayesian view of model complexity

Angelika van der Linde

FB03: Institute of Statistics, University of Bremen,
PO Box 330440, 28334 Bremen, Germany,
avdl@math.uni-bremen.de

Abstract

The paper addresses the problem of formally defining the ‘effective number of parameters’ in a Bayesian model which is assumed to be given by a sampling distribution and a prior distribution for the parameters. The problem occurs in the derivation of information criteria for model comparison which often trade off ‘goodness of fit’ and ‘model complexity’. It also arises in (frequentist) attempts to estimate the error variance in regression models with informative priors on the regression coefficients, for example in smoothing. It is argued that model complexity can be conceptualized as a feature of the joint distribution of the observed variables and the random parameters and might be formally described by a measure of dependence. The universal and accurate estimation of terms of model complexity is a challenging problem in practice. Several well-known criteria for model comparison are interpreted and discussed along these lines.

Key Words and Phrases: effective number of parameters, generalized degrees of freedom, information criteria, model comparison, mutual information

1 Introduction

“Unfortunately model order estimation remains a subject of tremendous controversy; there is little agreement on what the ‘best’ approach is, and indeed little agreement on if there is, in fact, such a thing as a ‘best’ approach.” (LANTERMAN, 2001, p. 186). Indeed, many ideas of model complexity have been around, from the ‘number of unknown parameters’ (AKAIKE, H. (1973)), ‘equivalent degrees of freedom’ (YE (1998), a measure of dependence between parameter estimates (BOZDOGAN (2010) and earlier papers cited therein), the posterior variance of the log-likelihood (WATANABE (2010) and earlier papers cited therein) to ‘description length’ or ‘coding length’ referring to coding theory (LANTERMAN (2001) and references therein). Formal definitions of model complexity are correspondingly dispersed.

Starting from the apparently dual nature in notions of model complexity which try to catch 'how much parameters and observations know about each other' and assuming a Bayesian approach where the parameter is random (initially endowed with a prior distribution), in this paper model complexity is conceptualized as a measure of stochastic dependence between observations and parameters. Without claiming that such a definition exactly comprises as special cases what has been discussed in the literature before, it is used as a benchmark. Typical situations that give rise to a measure of model complexity like smoothing in regression or the trade-off between 'model fit' and 'model complexity' in predictive model comparison are reviewed. In this way some major commonly used terms of model complexity can be identified as variants of the measure of dependence and explained by (i) different distributional assumptions for sampling (Gaussian, exponential families, general), (ii) the type of model comparison as prior or posterior predictive, (iii) the type of target in model comparison as expected utility with unknown true distribution or model specific with known model dependent distribution, (iv) the type of target in model comparison as representative or average. Thus not a single new definition but a way to think about it in information-theoretic terms is suggested.

The paper is organized as follows: In section 2 intuitions and heuristics about model complexity and resulting definitions are reviewed. In section 3 measures of dependence are introduced and elaborated for important special cases. In section 4 the occurrence of terms of model complexity in predictive model comparison is analyzed, and the effect of different set-ups on formal definitions is illustrated mainly for posterior predictive model comparison. In section 5 estimates of model comparison are summarized and related to different distributional assumptions. Section 6 concludes with a brief discussion of current developments.

2 Intuitions

2.1 What is a statistical model ?

Nearly everyone will agree that a statistical model describes the generating process of data $y = (y_1, ..., y_n)$ in terms of probability distributions or - for convenience - probability densities $p_0(y)$. The observations are certainly not independent, because in that case we couldn't learn from gathering more data. Hence it is assumed that their dependence is incorporated in a common parameter θ_0 of their probability distribution. In the simplest case $p(y|\theta_0) = \prod_{i=1}^{n} p(y_i|\theta_0)$, that is, the random variables Y_i are independent given θ_0. $p(y|\theta_0)$ is sometimes called a 'fully specified model', but I prefer to call it a 'fully specified conditional sampling distribution'. Even if we dare to specify it, it is not necessarily true, it is just a proposal hopefully compatible with the data. Usually we do not dare to specify θ_0 apriori, but propose a whole family of probability densities $\mathcal{P} = \{p(y|\theta)|\theta \in \Theta\}$ and then try to extract information from the data about θ. From a frequentist point of view and in earlier days of statistics $\mathcal{P}$ was often referred to as 'the model'. For example, a standard 'linear regression model' with covariables X_i and known regression functions $a(x_i)$ is specified by $Y_i = a(x_i)^T\beta + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ independent and identically distributed (iid). The maximum-likelihood estimator of (β, σ^2) can then be obtained without further assumptions. If however additional restrictions on β are imposed as for example $||\beta||^2 \leq$

τ^2 in ridge regression, in estimation such restrictions have to be taken into account, and the 'linear regression model' looks somewhat incomplete. In this spirit YE (1998) calls an estimator, a mapping $y \mapsto \widehat{\theta}(y) \in \Theta$, a 'modelling procedure'. Hence, in general, a conditional distributional assumption about the observations is to be completed by assumptions about the parameters. One principled way to do that is again in terms of probability distributions or prior densities. This is the set-up of Bayesian statistics which will be adopted here. In particular, a model M will be given by a family $\mathcal{P}$ of conditional sampling densities and a prior density $p(\theta)$ on the parameter space Θ. Inference and especially estimation will be based on the posterior density $p(\theta|y) = p(y|\theta)p(\theta)/p(y)$ according to Bayes' theorem. Although $M = (\mathcal{P}, p(\theta))$ is clearly an extension of just $\mathcal{P}$, a formal embedding of $\mathcal{P}$ into a model space $\mathcal{M} = \{(\mathcal{P}, p(\theta))\}$ requires the crucial definition of a 'non-informative' prior. Candidates are uniform priors, maximum entropy priors or reference priors. Inference at this borderline is often referred to as 'objective Bayesian' and is an active field of current research (see in particular the work of Berger and Bernardo, e.g. BERNARDO and SMITH (1994); BERNARDO (1997); BERNARDO, J. M. (2003); BERGER (2006)); BERGER et al. (2009)). Introducing a prior distribution, the parameter is conceptualized as a random variable ϑ with realizations θ, and a model M becomes equivalent to a joint density $p(y, \theta)$. It thus corresponds to a (tentative) representation of the density $p_0(y)$ by $E_\vartheta p(y|\vartheta)$, which is motivated and vindicated by de Finetti's representation theorem. The theorem, however, refers to the asymptotic empirical distribution function of Y to identify the generating process, whereas the proposed models $(\mathcal{P}, p(\theta))$ most often are approximate or simplified representations. In the process of learning the representations develop. In particular, having obtained data y, $(\mathcal{P}, p(\theta|y))$ is an adapted ('posterior') model. Although its dependence on data may look strange, an actual posterior is not qualitatively different from an informative prior. For example, conjugate priors are often interpreted as representing prior knowledge based on previous experiments or data. Hence from a Bayesian point of view to a family $\mathcal{P}$ sequences of models can be related as more and more data from replicated experiments become available.

2.2 What is model complexity ?

Here are some typical answers.

(i) 'Model complexity quantifies the explanatory power of θ for y, the potential of fitting y with θ.'

In a corresponding first attempt model complexity is linked to the number of parameters, for example p, the dimension of β in the linear regression. However, by restrictions on the parameters, in general an informative prior, model complexity as explanatory power is reduced, and the 'effective number of parameters' is expected to be smaller than p. Closely related is the notion of degrees of freedom (df) (of initially a χ^2-distribution), where in classical linear regression ($Y \sim N(\mu, \sigma^2 I_n)$, $\mu = A\beta$, $\beta \in \mathbb{R}^p$ without prior) $df = n - p$. With $E(Y) = \mu$ an unbiased estimate of σ^2 is $\widehat{\sigma}^2(y) = ||y - \widehat{\mu}(y)||^2/(n-p)$. To obtain an analogous estimate in smoothing, 'equivalent degrees of freedom' (edf) are required. WAHBA, 1990, ch. 5 suggested to define them by the trace of the 'hat matrix' S yielding the linear fit $\widehat{\mu}(y) = Sy$ of the data: $edf = n - tr(S)$. In classical linear regression the least squares estimate of β is

$\widehat{\beta}(y) = (A^TA)^{-1}A^Ty$ with $A = ((a_j(x_i)))_{j=1...p,i=1...n}$ and $a(x_i)^T = (a_1(x_i)...a_p(x_i))$. Hence $E(Y) = \mu = A\beta$ is estimated by $S = A(A^TA)^{-1}A^T$ and $tr(S) = p$. $tr(S)$ was widely accepted as measure of model complexity in regression (e.g. HASTIE and TIBSHIRANI, 1990, ch. 3.5; GREEN and SILVERMAN, 1994, ch. 3.3.4). Interestingly $tr(S)$ also illustrates a second idea about model complexity.

(ii) 'Model complexity quantifies the discriminatory power of y for θ', how hard it is to learn θ from data y, or how sensitive parameters are to perturbations of observations, or how large the estimation variance is. Again in classical linear regression (with $\theta = \beta$, σ^2 known) the estimation variance is measured by

$$tr(cov_{Y|\theta}(SY)) = \sigma^2 tr(S^2) \underset{S\,orth.\,projection}{=} \sigma^2 tr(S), \tag{1}$$

and the diagonal entries s_{ii} of S describe how sensitive $\widehat{\mu_i}(y)$ is to y_i. The resulting measure of sensitivity $tr(S) = \sum_i s_{ii}$ was generalized by YE (1998) and YE and WONG (1998) to a measure of model complexity called "general degrees of freedom of a modelling procedure". It is defined as

$$gdf(\theta) = \sum\nolimits_i cov_{Y|\theta}(\widehat{\mu_i}(Y), Y_i) = tr(cov_{Y|\theta}(\widehat{\mu}(Y), Y). \tag{2}$$

Without a prior θ is fixed as θ_0, marking the true but unknown distribution of Y assumed to belong to an exponential family, or θ is replaced by an estimate. In a similar spirit BOZDOGAN (2010) defines a measure of complexity on the covariance matrix of the parameter estimates. From a Bayesian point of view the notion of 'estimation variance' can be given two interpretations: one may refer to the reduction of uncertainty about θ by y (DEGROOT (1962)) or to the variability of the posterior distribution and related estimates of θ like the posterior mean $E(\vartheta|y)$. If uncertainty about θ is measured by $cov(\vartheta)$, the equation $cov(\vartheta) - E_Y(cov(\vartheta|Y)) = cov_Y(E(\vartheta|Y))$ shows that and how these two views are related.

(iii) Another, often rather implicit approach to model complexity is given by procedures of model comparison in terms of their predictive performance for future observations $\widetilde{y}$ generated in the same way as the data y. Although usually a good fit of the data by an estimate $\widehat{\theta}$ is required, too good a fit might result in a poor fit (prediction) of $\widetilde{y}$. Therefore, typically in predictive model comparison a trade-off between data fit and the sensitivity of $\widehat{\theta}(y)$ to y ('model complexity') is sought. The resulting information criteria (IC) all suggest some formal description of model complexity, as for example in

$$BIC = -2\log p(y|\widehat{\theta}_{ML}(y)) + 2p\log n \tag{3}$$

(SCHWARZ (1978)), in

$$AIC = -2\log p(y|\widehat{\theta}_{ML}(y)) + 2p \tag{4}$$

(AKAIKE, H. (1973)), or in

$$DIC = -2\log p(y|E(\vartheta|y)) + 2p_D \tag{5}$$

(SPIEGELHALTER et al. (2002)), where the measure of model complexity is $p_D = -2E_{\vartheta|y}(\log p(y|\vartheta)) - 2\log p(y|E(\vartheta|y))$ (in short: $p_D = \overline{D} - D(\overline{\theta})$ with D denoting deviance and $\overline{\theta}$ the posterior mean). p_D reduces to $tr(S)$ in linear regression. DIC might be seen as generalization of AIC in the sense that DIC reduces to AIC if the prior is non-informative. There are many more such criteria (e.g. GIC, KONISHI and KITAGAWA (1996); WAIC, WATANABE (2010) and hence unfortunately many formally different and often seemingly unrelated terms are claimed to be terms of model complexity. Technically they are derived setting up a measure of predictive success as a target (e.g. an average value of $\log p(\widetilde{y}|\widehat{\theta}(y)))$, and estimating that target by an empirical version (e.g. $\log p(y|\widehat{\theta}(y)))$. The average difference between the target and its estimate is thought to represent model complexity and is again estimated. This estimation can be involved, and often comprises simplifying approximations. The resulting terms are hardly comparable incorporating different targets and different approximation and estimation methods.

2.3 What is to be learnt ?

The brief review of ideas about model complexity points to several issues:

(i) Model complexity should be a dual concept, catching both directions of how much the parameter explains the observations and of how much the observations determine the parameter.

(ii) There is an ambiguity whether the second direction actually is from y to θ or rather from y to an estimate $\widehat{\theta}$. From a frequentist point of view the stochastic dependence between Y and $\widehat{\theta}(Y)$ is a natural measure to look at (as in (2)). From a Bayesian point of view ϑ is a random variable and the stochastic dependence between Y and ϑ is directly defined.

(iii) It is rather hopeless to derive a formal definition of model complexity from mathematical comparisons of the many versions representing it in the statistical literature on predictive model choice. Instead, we need to have an abstract notion explaining the many versions.

3 Measures of dependence between observables and parameters

The heuristics outlined in the previous section point to the idea of quantifying model complexity by a measure of dependence between observed random variables and parameters. Taking a Bayesian point of view (and thus interpreting frequentist inference as a special case with an 'objective' prior) measures of stochastic dependence between Y and ϑ are of interest, where as before their joint density is $p(y, \theta) = p(y|\theta)p(\theta) = p(\theta|y)p(y)$. A very general and often used measure is the mutual information defined by

$$I(Y, \vartheta) = E_{Y,\vartheta}[\log \frac{p(Y, \vartheta)}{p(Y)p(\vartheta)}]. \tag{6}$$

Recalling that the directed Kullback-Leibler divergence (KL-divergence) between two possible densities $p(z)$, $q(z)$ of a random variable Z is given by $KL(p, q) = E_{Z|p}[\log(p(Z)/q(Z))]$, the mutual information is seen to be the directed KL-divergence between the joint density $p(y, \theta)$ representing dependence and the product of marginal densities representing independence of Y and ϑ. A directed KL-divergence is not symmetric in the densities, $KL(p, q) \neq KL(q, p)$, but can be symmetrized adding the two directed divergences. Hence the symmetrized mutual information is given by

$$J(Y,\vartheta) = I(Y,\vartheta) + \widetilde{I}(Y,\vartheta), \tag{7}$$

where $\widetilde{I}(Y,\vartheta) = E_Y E_\vartheta(\log \frac{p(Y)p(\vartheta)}{p(Y,\vartheta)})$. These measures are invariant to re-parameterizations as is to be required for a measure of model complexity.

3.1 Properties and interpretations of (symmetrized) mutual information

3.1.1 Prior version

$I(Y,\vartheta)$ is symmetric in the variables Y and ϑ and thus has dual features. The decompositions $p(y,\theta) = p(y|\theta)p(\theta) = p(\theta|y)p(y)$ yield

$$I(Y,\vartheta) = E_\vartheta[KL(p(y|\vartheta), p(y))] = E_Y[KL(p(\theta|Y), p(\theta))]. \tag{8}$$

The middle term can be interpreted as a measure of variability of the (conditional) sampling density $p(y|\theta)$ relative to $p(y)$, catching the intuition of the 'modelling potential' of $p(y|\theta)$. The term on the right hand side measures the difference between prior and posterior distribution thus catching the intuition of 'difficulty of estimation' and 'sensitivity of an estimate' (the posterior density) to y. $I(Y,\vartheta)$ is a well studied quantity in Bayesian statistics, called "the expected amount of information about ϑ provided by an experiment yielding y" (e.g. GOEL (1983); BERNARDO and SMITH, 1994, p. 158). It is used as a starting point to derive 'objective' reference priors (BERGER et al. (2009)) or Bayesian experimental designs (e.g. CHALONER and VERDINELLI (1995)). Equation (8) carries over to $J(Y,\vartheta)$. By duality the mutual information also reflects restrictions of the (conditional) sampling distribution as related to a lack of identifiability or over-parametrization.

If $p(y|\theta)$ belongs to an exponential family, that is $p(y|\theta) = a(y)\exp(\theta^T t(y) - M(\theta))$, $J(Y,\vartheta)$ can be represented by the trace of a covariance matrix,

$$J(Y,\vartheta) = tr(cov_Y(E(\vartheta|Y), t(Y))) = tr(cov_\vartheta(\vartheta, E(t(Y)|\vartheta))), \tag{9}$$

and thus is similar to the generalized degrees of freedom in equation (2). But (2) and (9) are not equivalent, because gdf (as a frequentist concept) depends on θ, whereas $J(Y,\vartheta)$ is an integral over ϑ.

For example, in the special Gaussian case $Y|\theta \sim N(\theta,\Sigma)$ with known Σ, $t(y) = \Sigma^{-1}y$, and with the prior $\vartheta \sim N(0,K)$ one obtains for $y,\theta \in \mathbb{R}^p$

$$I(Y,\vartheta) = \frac{1}{2}\log(\det(I_p + K\Sigma^{-1})), \tag{10}$$

$$J(Y,\vartheta) = tr(K\Sigma^{-1}) = tr((I_p + K\Sigma^{-1})) - p. \tag{11}$$

If the prior becomes flat and hence intuitively the model becomes more complex, $I(Y,\vartheta)$ and $J(Y,\vartheta)$ increase. For example, if $\Sigma = \frac{\sigma^2}{n} I_p$ and $K = \tau^2 I_p$, $J(Y,\vartheta) = p(n\tau^2/\sigma^2)$. The posterior distribution is again Gaussian, $\vartheta|y \sim N(K(\Sigma+K)^{-1}y,$ $(\Sigma^{-1}+K^{-1})^{-1})$, and as $I_p+K\Sigma^{-1} = cov(\vartheta)cov(\vartheta|y)^{-1}$, mutual information compares the prior and posterior distribution in terms of the corresponding covariance matrices. $J(Y,\vartheta)$ is not equal to the trace of the 'hat matrix', though: for the 'hat matrix' in $E(\vartheta|y) = Sy$, $S = K(\Sigma+K)^{-1} = (\Sigma K^{-1}+I_p)^{-1}$ one has $tr(S) = tr(cov_{Y|\theta}(SY,\Sigma^{-1}Y) \neq tr(cov_Y(SY,\Sigma^{-1}Y) = J(Y,\vartheta)$.

3.1.2 Posterior version

The reasoning so far qualitatively corroborates the idea of formalizing model complexity by a measure of dependence between Y and ϑ, but $I(Y,\vartheta)$ and $J(Y,\vartheta)$ do not scale correctly if the number of parameters or the trace of the 'hat matrix' is taken as a benchmark. Looking at the problems yielding this benchmark, particularly AIC, reveals that in these set-ups it is not of interest how well the parameter θ fits the data, but how well an estimate of θ fits future observations $\widetilde{y}$. Hence the (symmetrized) mutual information between a random vector $\widetilde{Y}$ corresponding to future replications of the same experiment and the random variable ϑ_{post} corresponding to the posterior distribution may be more appropriate. The resulting joint density $\widetilde{p}(\widetilde{y},\theta) = p(y|\theta)p(\theta|y)$ with marginal densities $p(\widetilde{y}|y)$ and $p(\theta|y)$ yields $I(\widetilde{Y},\vartheta_{post})$ and $J(\widetilde{Y},\vartheta_{post})$ analogously to (6) and (7), which depend on the data y. This is coherent with the definition of a model by not only the family $\mathcal{P}$ =$\{p(y|\theta)|\theta\in\Theta\}$, but also a distribution on Θ, which evolves with replications of the experiment.

In the special Gaussian example

$$\begin{aligned} J(\widetilde{Y},\vartheta_{post}) &= tr(cov_{\vartheta_{post}}(\vartheta, E(\Sigma^{-1}\widetilde{Y})|\vartheta))) \\ &= tr(\Sigma^{-1}+K^{-1})^{-1}\Sigma^{-1} = tr((I_p+\Sigma K^{-1})^{-1}) \\ &= tr(S). \end{aligned} \tag{12}$$

If the prior becomes flat, $J(\widetilde{Y},\vartheta_{post})$ tends to $tr(I_p)=p$. For illustration, if again $\Sigma = \frac{\sigma^2}{n}I_p$ and $K=\tau^2 I_p$, $tr(S)\to p$ for $\tau^2\to\infty$ or $n\to\infty$. The Gaussian example is not only interesting in itself, its structure also lurks behind terms that occur if second order Taylor expansions are applied to approximate the sampling density, or if reference to asymptotic normality is made.

To conclude, some formulae to represent $J(\widetilde{Y},\vartheta_{post})$ in the non-Gaussian case are summarized and estimates are briefly discussed.

If $Y,\widetilde{Y}$ belong to an exponential family with $p(y|\theta) = a(y)\exp(\theta^T t(y) - M(\theta))$, the representation as in (9) holds,

$$J(\widetilde{Y},\vartheta_{post}) = tr(cov_{\vartheta_{post}}(\vartheta, E(t(\widetilde{Y})|\vartheta))). \tag{13}$$

Equation (13) is derived from

$$J(\widetilde{Y},\vartheta_{post}) = E_{\vartheta_{post}}E_{\widetilde{Y}|\vartheta}[(\vartheta-\overline{\theta})^T(t(\widetilde{Y})-t(y))], \tag{14}$$

where $\overline{\theta} = E(\vartheta|y)$. (14) also yields

$$J(\widetilde{Y},\vartheta_{post}) = E_{\vartheta_{post}}[KL(p(\widetilde{y}|\vartheta),p(\widetilde{y}|\overline{\theta})) + KL(p(\widetilde{y}|\overline{\theta}),p(\widetilde{y}|\vartheta))] \tag{15}$$

as demonstrated in (VAN DER LINDE (2004)).

The complexity term $p_D = \overline{D} - D(\overline{\theta})$ of DIC may be interpreted as estimate of $J(\widetilde{Y}, \vartheta_{post})$, if (15) holds and if $KL(p(\widetilde{y}|\theta), p(\widetilde{y}|\overline{\theta})) \approx KL(p(\widetilde{y}|\overline{\theta}), p(\widetilde{y}|\overline{\theta}))$:

$$\begin{aligned} & J(\widetilde{Y}, \vartheta_{post}) \\ = \ & E_{\vartheta_{post}}[KL(p(\widetilde{y}|\vartheta), p(\widetilde{y}|\overline{\theta})) + KL(p(\widetilde{y}|\overline{\theta}), p(\widetilde{y}|\vartheta))] \\ \approx \ & 2E_{\vartheta_{post}}[KL(p(\widetilde{y}|\vartheta), p(\widetilde{y}|\overline{\theta}))] \\ = \ & 2E_{\vartheta_{post}}[E_{\widetilde{Y}|\vartheta}(\log p(\widetilde{Y}|\vartheta) - \log p(\widetilde{Y}|\overline{\theta})] \\ \approx \ & 2E_{\vartheta_{post}}[\log p(y|\vartheta) - \log p(y|\overline{\theta})] \\ = \ & p_D \end{aligned}$$

(cp. SPIEGELHALTER et al., 2002, p. 604). Hence p_D can be expected to work as an estimate of $J(\widetilde{Y}, \vartheta)$ in exponential families. but not necessarily under general distributional assumptions.

By definition

$$J(\widetilde{Y}, \vartheta_{post}) = E_{\vartheta_{post}} E_{\widetilde{Y}|\vartheta}[\log p(\widetilde{Y}|\vartheta)] - E_{\vartheta_{post}} E_{\widetilde{Y}|y}[\log p(\widetilde{y}|\vartheta)].$$

Without reference to distributional assumptions PLUMMER (2002) suggested the representation

$$J(\widetilde{Y}, \vartheta_{post}) = E_{\vartheta^{(1)}_{post}} E_{\vartheta^{(2)}_{post}}[KL(p(\widetilde{y}|\vartheta^{(1)}), p(\widetilde{y}|\vartheta^{(2)}))] \tag{16}$$

on which Monte Carlo estimates can be based. For cross-validatory variants of this representation see PLUMMER (2008).

In general, estimation of (symmetrized) mutual information based on samples only, is a notoriously difficult problem because inherently unknown densities have to be estimated. In contrast, in model comparison the densities are specified, and this information considerably alleviates the estimation of model complexity.

3.2 Related ideas

Clearly other divergences between the joint density and the product of marginal densities could be used to define a measure of dependence. For example, the family of $\varphi-$ divergences between densities p and q of a random variable Z introduced by CSISZAR (1967),

$$D_\varphi(p, q) = E_q[\varphi(\frac{p(Z)}{q(Z)})]$$

where φ is continuous convex (and additionally satisfies non-restrictive regularity conditions) has been well studied (e.g. MICHEAS and ZOGRAFOS (2006) and references therein). The KL-divergence is obtained as a special case for $\varphi(u) = u\log(u)$. Features and comparisons of various measures of distance between probability distributions are currently discussed in the field of information geometry The (symmetrized) KL-divergence is closely related to fundamental statistical concepts like the principle of maximum likelihood and the notion of sufficiency, especially in exponential families (MCCULLOCH (1988)), and hence often represents the geometry of decision theory behind conventional statistical procedures. Therefore, although logically it is only one option, in terms of statistical practice it is an omnipresent and thus dominant option, often going unnoticed, though.

Similar in spirit to mutual information as measure of variability of $\log p(y|\theta)$ relative to $\log p(y) = \log E_\vartheta p(y|\vartheta)$, but formally different (because E_ϑ and $\log$ are interchanged), is the measure of model complexity introduced by WATANABE (2010). It is given by the sum of posterior variances of $\log p(y_i|\theta)$ (for conditionally independent observations y_i), which measures variability of $\log p(y_i|\theta)$ relative to $E_\vartheta(\log p(y_i|\vartheta))$.

The 'sensitivity of $p(y|\theta)$ to changes in θ' or dually the 'reduction of uncertainty about θ due to y' is often measured using the (expected) Fisher information matrix,

$$I(\theta) = ((E_{Y|\theta}[\frac{\partial \log p(Y|\theta)}{\partial \theta_i}\frac{\partial \log p(Y|\theta)}{\partial \theta_j}]))_{i,j=1\ldots p}. \tag{17}$$

In differential geometry a family of sampling distributions corresponds to a statistical manifold, and the Fisher information characterizes the curvature of the log-likelihood functions. (See MURRAY and RICE (1993) or AMARI et al. (1987) Thus Fisher information describes the magnitude of change locally. $I(\theta)$ provides an approximation

$$KL(p(y|\theta), p(y|\theta')) \approx \frac{1}{2}(\theta - \theta')^T I(\theta)(\theta - \theta'), \tag{18}$$

(BLYTH (1994)). This is different from what is intended to grasp in a description of model complexity by the global range of densities. In general, $I(Y, \vartheta)$ or $J(Y, \vartheta)$ cannot be expressed as an expected KL-divergence of two densities in the same family $\mathcal{P}$ because in mutual information the reference density is the marginal density which need not belong to $\mathcal{P}$. $p(y)$ can only replaced by $p(y|\theta')$ for some θ' in special cases, for example by $p(y|E(\vartheta))$ in $J(Y, \vartheta)$, if $\mathcal{P}$ is an exponential family. (BOZDOGAN (2010) and earlier papers cited therein) introduces a measure of model complexity based on the covariance matrix of parameter estimates, especially based on the inverse Fisher information matrix referring to the asymptotic distribution of maximum likelihood estimates.

4 Model complexity in predictive model comparison

Information criteria for predictive model comparison very often trade off 'model fit' and 'model complexity'. It is useful to distinguish between criteria assessing 'prior prediction', that is, accommodating observations y using the prior $p(\theta)$, or 'posterior prediction', that is, predicting future observations $\widetilde{y}$ (of the same type as y) using the posterior $p(\theta|y)$. In the former case, for example the marginal likelihood $\log p(y) = \log(E_\vartheta[p(y|\vartheta)])$ is of interest, in the latter case for example $E_{\widetilde{Y}|y}[\log p(\widetilde{Y}|y)] = E_{\widetilde{Y}|y}\log(E_{\vartheta_{post}}[p(\widetilde{Y}|\vartheta)])$ or $E_{\widetilde{Y}}E_{\vartheta_{post}}[\log p(\widetilde{Y}|\vartheta)]$ is to be considered. $\widetilde{y}$ is a replicate vector, the result of running again the experiment yielding y. No assumption of independence (conditionally on θ) of the components in $Y, \widetilde{Y}$ is made at this stage.

The key idea to be discussed in this section is that the decomposition of such criteria into terms of 'model fit' and 'model complexity' is due to a fundamental representation of marginal entropy. For two random variables U and V with joint probability density $p(u,v)$ one has: “Marginal entropy equals conditional entropy plus mutual information”. Formally, based on

$$-\log p(u) = -\log p(u|v) + \log(\frac{p(u,v)}{p(u)p(v)}),$$

one obtains, taking expectations with respect to (U,V),

$$E_U[-\log p(U)] = E_V E_{U|V}[-\log p(U|V)] + E_{U,V}[\log(\frac{p(U,V)}{p(U)p(V)})], \tag{19}$$

or in short,

$$H(U) = H(U|V) + I(U,V). \tag{20}$$

Similarly,

$$E_U E_V[-\log p(U|V)] = H(U|V) + J(U,V). \tag{21}$$

There are prior versions of (20), (21) corresponding to $U = Y,\ V = \vartheta$ and posterior versions corresponding to $U = \widetilde{Y},\ V = \vartheta_{post}$. It is proposed that many information criteria occur as variants of these core equations without claiming that they are all just special cases. The ‘fluctuations’ of popular information criteria around (20) and (21) reflect options and choices that are made with respect to the target (on the left hand side) and the estimation of approximate targets (on the right hand side). Some of these choices, especially for posterior predictive model comparison, are briefly discussed. Typically, referring to an information criterion only the (estimated, approximate) right hand side is quoted. In order to gain insight, however, the derivation from the target has to be tracked.

4.1 Targets

4.1.1 Expected utilities versus model specific targets

Often in predictive model comparison the utility $u(M,z)$ of a model M is described by a predicting density $\widetilde{p}_M$ evaluated at an observation z.

In ‘prior prediction’ typically the marginal density evaluated at the data is used, $u(M,y) = \log p_M(y)$, and the model maximizing this term is chosen. If there are only finitely many models under consideration, and a uniform prior is specified over $\mathcal{M}$, $p_M(y)$ is proportional to the posterior probability of that model. On average over possible data sets generated according to M the approach corresponds to minimizing the entropy of the model specific marginal density, and (20) indicates that $I(Y,\vartheta)$ is the corresponding term of model complexity.

There are other ideas about model complexity in prior prediction derived from coding theory as the criteria of minimum description length (MDL) or minimum message length (MML). These ideas and those arising from statistics like the ‘Bayesian information criterion’ (BIC) can - at least in parts - be formally related within the field of information theory (LANTERMAN (2001)). These links are not explored in this paper and require further investigation.

In ‘posterior prediction’ the approach is involved. The utility of a model is now described by a predicting density evaluated in future observations $\widetilde{y}$, for example, $u(M,\widetilde{y}) = \log p_M(\widetilde{y}|y)]$

or $u(M,\widetilde{y}) = E_{\vartheta_{post}}[\log p_M(\widetilde{y}|\vartheta)]$, or particularly from a frequentist point of view $u(M,\widetilde{y}) = \log p_M(\widetilde{y}|\widehat{\theta}(y))$. As the true (marginal) density p_0 of Y and hence of $\widetilde{Y}$ as observation of the same type is unknown, the crucial point then is to evaluate an average over future observations, an expected utility $E_{\widetilde{Y}}[u(M,\widetilde{Y})]$. Frequentists and Bayesians might agree on the existence of a true density p_0 as being asymptotically determined (for Bayesians giving rise to de Finetti's theorem), but it is not available and all competing models only suggest possible approximations. The evaluation of a target specified as expected utility therefore requires choosing a common density $\widehat{p}_0$ for $\widetilde{Y}$. Major choices are: a density corresponding to a more complex model than those under competition like a model average (e.g. SAN MARTINI and SPEZZAFERI (1984)), an encompassing model (e.g. GUTIERREZ-PENA, E. (1987)), a nonparametric model (e.g. GUTIERREZ-PENA and WALKER (2001)), or - if the components of $Y, \widetilde{Y}$ are independent and identically distributed (iid) - the empirical distribution function given the data (e.g. KONISHI and KITAGAWA (1996), GIC; ANDO (2007), BPIC). The independence assumption also leads to cross-validatory targets (e.g. PLUMMER (2008). The key question here is: how can an expected utility with respect to a common density of $\widetilde{Y}$ result in a decomposition into model specific terms of complexity ?

For illustration, remember that the expected utility yielding AIC,
$-2E_{\widetilde{Y}|\theta_0}E_{Y|\theta_0}[\log p_M(\widetilde{Y}|\widehat{\theta}_{ML}(Y))]$, and the model specific target of DIC,
$-2E_{\widetilde{Y}|y}[\log p_M(\widetilde{Y}|E(\vartheta|y))]$, for a non-informative prior both result in the criterion
$-2\log p_M(y|\widehat{\theta}_{ML}(y)) + 2p$. Similarly EFRON (1986) starts with a model specific target
$-2E_{\widetilde{Y}|\theta}[\log p_M(\widetilde{Y}|\widehat{\theta}_{ML}(y))]$ with $p_M(y|\theta)$ in an exponential family, and derives
$C_M(\theta) = tr(cov_{\widetilde{Y}|\theta}(\widehat{\theta}_{ML}(\widetilde{Y}), \widetilde{Y}))$ as a complexity term (cp. equations (2) and (9)) and ends up again with $2p$ as approximation of $C_M(\theta)$ for all θ as in AIC.

The key question is answered in three parts. (In the sequel the subscript for the model will be dropped again.)

- In general, under an expected utility the complexity term is not equal to $I(\widetilde{Y}, \vartheta_{post})$ or $J(\widetilde{Y}, \vartheta_{post})$. For example, analogously to (21), with $U = \widetilde{Y}$, $V = \vartheta_{post}$, but expected utility with respect to p_0

$$-E_{\widetilde{Y}|p_0}E_{\vartheta_{post}}[\log p(\widetilde{Y}|\vartheta)] = -E_{\widetilde{Y}|p_0}E_{\vartheta_{post}|\widetilde{y}}[\log p(\widetilde{Y}|\vartheta)] + E_{\widetilde{Y}|p_0}[J(\widetilde{Y})],$$

  and $J(\widetilde{Y}, \vartheta_{post}) = E_{\widetilde{Y}|y}[J(\widetilde{Y})] \neq E_{\widetilde{Y}|p_0}[J(\widetilde{Y})]$.

- Under a 'good model assumption' $p(\widetilde{y}|y) \approx p_0(\widetilde{y})$, the complexity terms $E_{\widetilde{Y}|y}[J(\widetilde{Y})]$ and $E_{\widetilde{Y}|p_0}[J(\widetilde{Y})]$ are close. Under the Bayesian paradigm of sequential learning $p(\widetilde{y}|y)$ represents the current belief, what is known about $\widetilde{Y}$ to the best of our present knowledge.

- Second order Taylor expansions of $\log p(\widetilde{y}|\theta)$ induce Gaussian approximations, for which under a non - informative prior $J(\widetilde{Y}, \vartheta_{post}) \approx p$, and hence the complexity terms coincide for all initially different $\widetilde{Y} \sim N(\theta, \Sigma_M)$ with model specific covariance matrix Σ_M. This is the special case in recovering AIC in spite of starting with a model specific target (DIC or EFRON (1986)).

Of course, there are also model specific targets set-up for posterior predictive model comparison. For example, the target of DIC (SPIEGELHALTER et al. (2002)), its cross-validatory version (PLUMMER (2008)), the criteria by GELFAND and GHOSH (1998). In these cases $J(\widetilde{Y}, \vartheta_{post})$ can be more immediately related to the term of model complexity occurring in the criteria.

The ICOMP-type criteria by (BOZDOGAN (2010) and earlier references therein) which are intended to generalize AIC do not fit into this framework being based on utilities different from those specified in (20) and (21). The same applies to targets that do not take into account the full sampling distribution, as - for example - averaged squared errors $E_{Y|\mu}(||\mu - \widehat{\mu}(Y)||^2)$, popular in smoothing and regression with linear estimators $\widehat{\mu}(Y)$. These are not discussed here.

4.1.2 Representative versus average targets

A density based target is called representative if an estimate is plugged in as, for example, in $E_{\widetilde{Y}|y}[\log p(\widetilde{Y}|\widehat{\theta}(y))]$. If, in contrast, θ is integrated out, as, for example, in $E_{\widetilde{Y}} E_{\vartheta_{post}}[\log p(\widetilde{Y}|\vartheta)]$, it is called an average (density based) target. Representative targets are necessary from a traditional frequentist point of view where no prior is assumed. Representative targets obviously depend on the estimator $\widehat{\theta}$, which should be invariant under re-parameterizations. Average targets are natural from a Bayesian point of view though representative targets as expected utilities are not excluded. Strictly speaking (20) and (21) apply only to average targets, but analogous expressions can be derived for representative targets. For example,

$$-E_{\widetilde{Y}|y}[\log p(\widetilde{Y}|\widehat{\theta}(y))] = H(\widetilde{Y}|\vartheta_{post)} + E_{\vartheta_{post}}[KL(p(\widetilde{y}|\vartheta), p(\widetilde{y}|\widehat{\theta}(y)))]. \tag{22}$$

Recalling that $I(\widetilde{Y}, \vartheta_{post}) = E_{\vartheta_{post}} KL(p(\widetilde{y}|\vartheta), p(\widetilde{y}|y))$ and the interpretation of Kullback-Leibler divergences as measures of variability of $p(\widetilde{y}|\theta)$ with respect to $p(\widetilde{y}|y)$, it is seen that essentially for representative targets the density $p(\widetilde{y}|\widehat{\theta}(y))$ rather than the marginal $p(\widetilde{y}|y)$ is used as a reference. In order to formally link mutual information and Kullback-Leibler divergences incorporating estimates a representation

$$J(\widetilde{Y}, \vartheta_{post}) = E_{\vartheta_{post}}[KL(p(\widetilde{y}|\vartheta), p(\widetilde{y}|\widehat{\theta}^*(y)))] + E_{\vartheta_{post}}[KL(p(\widetilde{y}|\widehat{\theta}^*(y)), p(\widetilde{y}|\vartheta))] \tag{23}$$

would be required for some estimator $\widehat{\theta}^*$. Such a representation indeed holds for $J(\widetilde{Y}, \vartheta_{post})$ if $p(y|\theta)$ belongs to an exponential family and $\widehat{\theta}^*(y) = E(\vartheta|y) = \overline{\theta}$ (as outlined in section 3.1.2). Furthermore under these conditions, within a second order Taylor expansion

$$\frac{1}{2} J(\widetilde{Y}, \vartheta_{post}) \approx E_{\vartheta_{post}}[KL(p(\widetilde{y}|\vartheta), p(\widetilde{y}|\overline{\theta}))] \approx E_{\vartheta_{post}}[KL(p(\widetilde{y}|\overline{\theta}), p(\widetilde{y}|\vartheta))], \tag{24}$$

(Kullback, 1968, ch. 6.2), such that (22) becomes

$$-E_{\widetilde{Y}|y}[\log p(\widetilde{Y}|\widehat{\theta}(y))] \approx H(\widetilde{Y}|\vartheta_{post}) + \frac{1}{2} J(\widetilde{Y}, \vartheta_{post}). \tag{25}$$

Hence, if $p(y|\theta)$ belongs to an exponential family, the model complexity corresponding to a representative target can be approximately described again in terms of $J(\widetilde{Y}, \vartheta_{post})$, however with a scaling factor of $1/2$. This result explains many criteria that have been proposed in the literature. Without restrictive distributional assumptions $J(\widetilde{Y}, \vartheta_{post})$ is not the term of model complexity associated with a representative target. The right hand side of (24) corresponding to a term of model complexity for a representative target (with representer $\overline{\theta}$), may well be estimated by $p_D/2$ beyond exponential families under the assumption of symmetry as argued in section 3.1.2.

4.2 Estimation of targets

4.2.1 Model complexity as penalty for using the data twice

In the literature about posterior predictive model comparison it is argued that model complexity results from the "bias correction for using the data twice".

Introducing a direct estimate $\widehat{T}(y, \widehat{\theta}(y))$ of a representative target $T(\widehat{\theta}(y))$ is a traditional main steam of reasoning in posterior predictive model comparison, yielding the famous decomposition into 'model fit' $(= \widehat{T}(y, \widehat{\theta}(y)))$ and 'model complexity' $(\doteq T(\widehat{\theta}(y)) - \widehat{T}(y, \widehat{\theta}(y)))$ in information criteria.
For example, in the derivation of DIC with the target $T(\overline{\theta}) = -2E_{\widetilde{Y}|y}[\log p(\widetilde{Y}|\overline{\theta})]$, $\widehat{T}(y, \overline{\theta}) = -2\log p(y|\overline{\theta}) = D(\overline{\theta})$ one obtains $T(\overline{\theta}) - \widehat{T}(y, \overline{\theta}) \approx 2p_D = 2(\overline{D} - D(\overline{\theta}))$ (SPIEGELHALTER et al., 2002, p. 604). If (24) holds, in particular for $\widehat{\theta}(y) = \overline{\theta}$ in DIC,
$p_D \approx 2E_{\vartheta_{post}}[KL(p(\widetilde{y}|\vartheta), p(\widetilde{y}|\overline{\theta}))]$ estimates $J(\widetilde{Y}, \vartheta_{post})$. Note that, in comparison to (21), $\widehat{T}(y, \widehat{\theta}(y))$ does not estimate $2H(\widetilde{Y}|\vartheta_{post})$. For example, in DIC, $\widehat{T}(y, \widehat{\theta}(y)) = D(\overline{\theta})$, whereas $2H(\widetilde{Y}|\vartheta_{post})$ is estimated by $\widehat{T}(y) = \overline{D}$. Hence the decomposition $DIC = \overline{D} + p_D$ is one into 'model adequacy plus model complexity' and corresponds to (21), whereas the decomposition $DIC = D(\overline{\theta}) + 2p_D$ is one into 'fit' plus two times 'model complexity'.

As seen in the previous section (comparing (21) to (22)) the transition from an average target like $T = -2E_{\widetilde{Y}|y}E_{\vartheta_{post}}[\log p(\widetilde{Y}|\vartheta)]$ to the representative target $T(\widehat{\theta}(y)) = -2E_{\widetilde{Y}|y}[\log p(\widetilde{Y}|\widehat{\theta}(y))]$ 'costs' $2J(\widetilde{Y}, \vartheta_{post}) -$
$2E_{\vartheta_{post}}[KL(p(\widetilde{y}|\vartheta), p(\widetilde{y}|\widehat{\theta}(y)))]$, and (under the conditions of exponential families and symmetry) the last term is close to $2E_{\vartheta_{post}}[KL(p(\widetilde{y}|\vartheta), p(\widetilde{y}|\overline{\theta}))] = J(\widetilde{Y}, \vartheta_{post})$. Combining the two steps the overall difference $T - \widehat{T}(y, \widehat{\theta}(y))$ for an average target becomes $3J(\widetilde{Y}, \vartheta_{post})$.

A more general argument particularly for an average target without reference to a representative target, that using the data twice yields model complexity in terms of $J(\widetilde{Y}, \vartheta_{post})$ has not yet been presented. The two conditions required in the reasoning above are ubiquitous in the literature: exponential families are common in regression analyses, a major field where model comparison is required, and symmetry is achieved by second order Taylor expansions.

A plug-in estimate $\widehat{T}$ of a target T and an analytical approximation of the difference is not always necessary, though. Simulation based estimates of the target or the 'bias' might do as well. Furthermore, if model complexity is assessed as 'bias' applying analytical expansions, approximations and simplifications to $T - \widehat{T}$, do these perform uniformly well across the candidate models ? Or do model specific approximation errors distort the model comparison ? If so, do sampling based methods provide uniformly better estimates ?

4.2.2 Simulation based estimates

Simulation based estimates are obtainable if the target is model specific or, for targets that are expected utilities, if the observations are independent conditionally on θ. Up to now this assumption has not been made. The following remarks do not provide additional insight into model complexity but are included to complete the review of main ideas in predictive model comparison.

Model specific targets In prior predictive model comparison the quantity of interest is the 'model evidence' $E_\vartheta[p(y|\vartheta)]$. An influential paper on (Gibbs) sampling of values of a marginal density was that by CHIB (1995), a more recent paper is (NEAL (2001)). New techniques are described by DIDELOT et al. (2011), who also give brief reviews and provide many references.

In posterior predictive model comparison the problem of Monte Carlo estimation of model specific targets like that of (21) or GELFAND and GHOSH (1998) has hardly been tackled. This maybe due to the fact that model specific targets are not that frequent, the dominating approach being that of expected utilities. Here is a field of research opportunities.

Estimates of expected utility under independence assumptions The assumption that, conditionally on θ, the observations are independent yields

$$\log p(\widetilde{y}|\theta) = \sum\nolimits_i \log p(\widetilde{y}_i|\theta). \tag{26}$$

If the random variables Y_i are iid, the random variables $\widetilde{Y}_i$ are iid, too. If the random variables Y_i and hence $\widetilde{Y}_i$ correspond to covariables X_i as in regression, $\log p(\widetilde{y}_i|\theta) = \log p(\widetilde{y}_i|x_i, \theta)$, and (26) can have two meanings: (i) the pairs (X_i, Y_i) are independent, or (ii) given the x_i the Y_i are independent. In case (i) the replication of the experiment may result in new values of the covariables, in case (ii) the experimental design is fixed.

Under the independence assumption an empirical distribution function for (X, Y) or Y can be obtained. If the target is an expected utility the empirical distribution function can be used to replace the unknown distribution function in the approximation of the 'bias' (e.g. KONISHI and KITAGAWA (1996); ANDO (2007)). It can also be used to derive bootstrap estimates of the 'bias' (e.g. SHIBATA (1997)).

Alternatively a cross-validatory substitute of a target can be obtained based on

$$T = E_{\widetilde{Y}_i}[\log p(\widetilde{Y}_i|\widehat{\theta}(y))] \cong \frac{1}{n}\sum\nolimits_i \log p(y_i|\widehat{\theta}(y_{-i})) = T_{CV}$$

for a representative target or on

$$T = E_{\vartheta_{post}}[\log p(\widetilde{y}_i|\vartheta)] \cong \frac{1}{n}\sum\nolimits_i E_{\vartheta_{post-i}}[\log p(y_i|\vartheta)] = T_{CV}$$

for an average target, where y_{-i} denotes the vector of observations without y_i and $E_{\vartheta_{post-i}}$ the posterior expectation with respect to y_{-i}. The approximations T_{CV} can be evaluated directly, which might be computationally expensive, or they can again be estimated using the data twice, which then requires a correction term. For example,

$$T_{CV} = \frac{1}{n}\sum\nolimits_i E_{\vartheta_{post-i}}[\log p(y_i|\vartheta)] \approx \frac{1}{n}\sum\nolimits_i E_{\vartheta_{post}}[\log p(y_i|\vartheta)] := \widehat{T}_{CV}.$$

(For details and applications see PLUMMER (2008).) The correction term $T_{CV} - \widehat{T}_{CV}$ corresponds to $\frac{1}{n}\sum_i J(Y_i, \vartheta_{post-i})$, but not to $\frac{1}{n}\sum_i J(\widetilde{Y}_i, \vartheta_{post})$ as would be expected under independence with $\frac{1}{n}J(\widetilde{Y}, \vartheta_{post})$ as term of model complexity. The difference is due to the cross-validatory approximation of the target $T \approx T_{CV}$ which 'reduces the sample size to $n-1$' (cp. EFRON (1986)). Asymptotically the information criterion and cross-validation are equivalent (STONE (1977) for AIC; WATANABE (2010) for WAIC). Watanabe's target, the expected utility $T = -E_{\widetilde{Y}}[\log p(\widetilde{Y}|y)]$ is replaced by $T_{CV} = -\frac{1}{n}\sum_i \log p(y_i|y_{-i})$, which in turn is estimated by $\widehat{T}_{CV} = -\frac{1}{n}\sum_i \log p(y_i|y)$ yielding

$$WAIC = -\frac{1}{n}\sum\nolimits_i \log p(y_i|y) + \sum\nolimits_i var_{\vartheta_{post}}[\log p(y_i|\vartheta)]. \tag{27}$$

5 Discussion

The brief review has revealed a few crucial and perhaps controversial issues in predictive model comparison that are to be summarized here.

1. Joint features and differences of frequentist and Bayesian predictive model comparison become aware only gradually. In traditional frequentist statistics the sampling distribution of parameter estimates and the (inverse) Fisher information matrix characterizing the asymptotic distribution of maximum likelihood estimates have been prevailing concepts. The Bayesian approach not only is more comprehensive but also results in emphasis on different tools.

- The focus is shifted from representative to average targets in predictive model comparison. Representative targets most often reflect the frequentist tradition in statistics. From a Bayesian point of view representative targets may be set up, but do not take into account uncertainty about θ, average targets seem more appropriate.

- Ubiquitous second order Taylor expansions are recognized to be related to representative targets, while average targets more naturally combine with information theoretic decompositions as in (20) and (21). Average targets correspond to mutual information as measure of model complexity, representative targets correspond to KL-divergences with a representative (rather than marginal) density as reference density. These in general different measures of model complexity coincide in exponential families for the symmetrized KL-divergence with the mean of ϑ as representer.

- Experience with simulation based methods of estimation of targets is growing. They have already been applied to prior predictive targets but less so to (model specific) posterior predictive targets. Under independence, bootstrapping is still an option.

- Comparisons based on finite sample sizes can and should be made. Asymptotic equivalences between information criteria and cross-validatory procedures demonstrate that eventually the same structures of targets are caught. In practice, however, the preference or disadvantage of a model due to an approximation or finite sample estimation is of interest.

2. A common language for comparing stochastic models is the language of probability theory. Basis concepts of probability theory are entropy and information. “Many problems of current scientific interest are described, at least colloquially, as being about the flow and control of information. It has been more than fifty years since Shannon formalized our intuitive notions of information, yet relatively few fields actually use information theoretic methods for the analysis of experimental data” (SLONIM et al., 2005, p.1). Information theoretic concepts have been lurking behind much of decision theory applied in statistics. Notwithstanding context - or subject-specific aims of some statistical analyses, information theory provides a meta-level to formalize general interests in data analysis. It often acts as a ‘spirit’ driving the scientists’ intuitions. For instance, SPRENGER (2010) discussing the “weight of simplicity in statistical model comparison” asks if there is a definite trade-off between simplicity and goodness of fit. Although no particular criterion provides an answer, a “normative force” can be recognized in the decompositions (20) and (21).

An information-theoretic interpretation of the notion of model complexity was proposed in this paper, intended to invite and stimulate further research and discussion.

3. The discussion emphasized that it s useful to study information criteria as estimates of targets. A crucial step of approximation/estimation in such a derivation is the good model assumption. In particular, if If an independence assumption does not hold and hence a cross-validatory target cannot be defined, is it acceptable and even more honest to define a model specific target rather than invoking some ‘good model assumption’ hidden in the evaluation of approximate terms ? From a Bayesian point of view in prior prediction model specific targets might be justified as proportional to posterior probabilities of models under a uniform prior (on models), if finitely many models are to be compared. How are they justified in posterior prediction ?

Another crucial issue is the appropriateness of approximations under general distributional assumptions. What can be said about the uniformity of performance of an information criterion as estimator of a target across models under comparison ? Reference to the decompositions (20) and (21) may provide a benchmark. For example, in information criteria, knowing the quantity to be estimated reveals limitations of frequently used estimates: p, the number of parameters is a bad estimator, if Normality does not hold or the prior is informative; p_D might become a bad estimator of $J(\widetilde{Y}, \vartheta_{post})$ outside exponential families.

4. “Singular models”, where commonly assumed regularity conditions (like positive definiteness of the Fisher information matrix) do not hold, are more frequently used, and the famous information criteria developed for regular models like AIC or BIC need to be generalized. WATANABE (2010) throughout his work points especially to neural networks, mixtures, reduced rank regression, hidden Markov models etc., which have come into use with the ability to evaluate posterior distributions by sampling. Along with the availability of powerful computing resources, statistical methods have been re-thought and advanced also in the machine learning community, and any attempt to extract principles of scientific reasoning from statistical practice

has to take that work into account as well.

Acknowledgement
I thank an anonymous referee and Jan Sprenger (Tilburg) for comments that helped to improve an earlier draft of this paper.

References

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In B. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281. Akademiai Kaido, Budapest, 1973.

S. AMARI, O. E. BARNDORFF-NIELSEN, R. E. KASS, S: L. LAURITZEN, and C. R. RAO. *Differential Geometry in Statistical Inference*. Springer, Berlin, 1987.

T. ANDO. Bayesian predictive information criterion for the evaluation of hierarchical bayesian and empirical bayesian models. *Biometrika*, 94:443–458, 2007.

J. O. BERGER. The case for objective bayesian analysis. *Bayesian Analysis*, 1:385–464, 2006. (with discussion).

J. O. BERGER, J. M. BERNARDO, and D. SUN. The formal definition of reference priors. *Annals of Statistics*, 37:905–938, 2009.

J. M. BERNARDO. Noninformative priors do not exist. *Journal of Statistical Planning and Inference*, 65:159–189, 1997. (with discussion).

J. M. BERNARDO and A. F. M. SMITH. *Bayesian Theory*. Wiley, Chichester, 1994.

BERNARDO, J. M. Bayesian statistics. In R. VIERTL, editor, *Encyclopedia of Life Support Systems (EOLLS)*. UNESCO, Oxford, 2003.

S. BLYTH. Local divergence and association. *Biometrika*, 81:579–584, 1994.

H. BOZDOGAN. A new class of information complexity (icomp) criteria with an application to customer profiling and segmentation. *Istanbul University Journal of the School of Business Administration*, 39:370–398, 2010.

K. CHALONER and I. VERDINELLI. Bayesian experimental design. a review. *Statistical Science*, 10:273–304, 1995.

S. CHIB. Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, 90:1313–1321, 1995.

I. CSISZAR. Information type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

M. H. DEGROOT. Uncertainty, information and sequential experiments. *Annals of Mathematical Statistics*, 33:404–419, 1962.

X. DIDELOT, R. G. EVERITT, A. M. JOHANSEN, and D. J. LAWSON. Likelihood-free estimation of model evidence. *Bayesian Analysis*, 6:1–30, 2011.

B. EFRON. How biased is the apparent error rate of a prediction rule ? *Journal of the American Statistical Association*, 74:461–470, 1986.

A. E. GELFAND and S. K. GHOSH. Model choice: A minimum posterior predictive loss approach. *Biometrika*, 85:1–11, 1998.

P. GOEL. Information measures and bayesian hierarchical models. *Journal of the American Statistical Association*, 78:408–410, 1983.

P. J. GREEN and B.W. SILVERMAN. *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, London, 1994.

E. GUTIERREZ-PENA and S. WALKER. A bayesian predictive approach to model selection. *Journal of Statistical Planning and Inference*, 93:259–276, 2001.

GUTIERREZ-PENA, E. A bayesian predictive semiparametric approach to variable selection and model comparison in regression. In *Proceedings of the 51st session of the ISI*, pages 17–29, Istanbul, 1987.

T. J. HASTIE and R.J. TIBSHIRANI. *Generalized Additive Models*. Chapman and Hall, London, 1990.

S. KONISHI and G. KITAGAWA. Generalised information criteria in model selection. *Biometrika*, 83:875–890, 1996.

S. Kullback. *Information Theory and Statistics*. Dover, Mineola, New York, 1968.

A. D. LANTERMAN. Schwarz, wallace and rissanen: Intertwining themes in theories of model selection. *International Statistical Review*, 69:185–212, 2001.

R. E. MCCULLOCH. Information and the likelihood function in exponential families. *The American Statistician*, 42:73–75, 1988.

A. C. MICHEAS and K. ZOGRAFOS. Measuring stochastic dependence using $\varphi-$ divergence. *Multivariate Analysis*, 97:765–784, 2006.

A. C. MURRAY and J. W. RICE. *Differential Geometry and Statistics*. Chapman and Hall, London, 1993.

R. NEAL. Annealed importance sampling. *Statistics and Computing*, 11:125–139, 2001.

M. PLUMMER. Discussion of ’bayesian measures of model complexity and fit’. *Journal of the Royal Statistical Society B*, 64:620–621, 2002.

M. PLUMMER. Penalized loss functions for bayesian model comparison. *Biostatistics*, 9:523–539, 2008.

A. SAN MARTINI and F. SPEZZAFERI. A predictive model selection criterion. *Journal of the Royal Statistical Society B*, 46:296–303, 1984.

G. SCHWARZ. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

R. SHIBATA. Bootstrap estimate of kullback-leibler information for model selection. *Statistica Sinica*, 7:375–394, 1997.

N. SLONIM, G. S. ATWAL, G. TKACIK, and W. BIALEK. Estimating mutual information and multi-information in large networks. arXiv:cs.IT/0502017v1, 2005.

D. SPIEGELHALTER, N. BEST, B. P. CARLIN, and A. VAN DER LINDE. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, 64:583–639, 2002.

J. SPRENGER. The weight of simplicity in statistical model comparison. Technical report, Center for Logic and Philosophy of Science, Tilburg University, Netherlands, 2010. URL:www.laeuferpaar.de/papers.html.

M. STONE. An asymptotic equivalence of choice of model cross-validation and akaike’s criterion. *Journal of the Royal Statistical Society B*, 36:44–47, 1977.

A. VAN DER LINDE. On the association between a random parameter and an observable. *Test*, 13:85–111, 2004.

G. WAHBA. *Spline Models for Observational Data*. SIAM, Philadelphia, Pennsylvania, 1990.

S. WATANABE. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 1:1–48, 2010.

J. YE. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93:120–131, 1998.

J. YE and W. WONG. Evaluation of highly complex modeling procedures with binomial and poisson data. Technical report, Graduate School of Business, University of Chicago, 1998.